BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 1 2007, pages 5–13
doi:10.1093/bioinformatics/btl549

Sequence analysis

Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays

Waibhav Tembe, Nela Zavaljevski, Elizabeth Bode[1], Catherine Chase[1], Jeanne Geyer[1], Leonard Wasieloski[1], Gary Benson[2] and Jaques Reifman*

Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Ft. Detrick, MD, [1]Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases, Ft. Detrick, MD and [2]Departments of Biology and Computer Science, Boston University, Boston, MA, USA

Received on June 15, 2006; revised on October 18, 2006; accepted on October 21, 2006
Advance Access publication October 26, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Advances in DNA microarray technology and computational methods have unlocked new opportunities to identify 'DNA fingerprints', i.e. oligonucleotide sequences that uniquely identify a specific genome. We present an integrated approach for the computational identification of DNA fingerprints for design of microarray-based pathogen diagnostic assays. We provide a quantifiable definition of a DNA fingerprint stated both from a computational as well as an experimental point of view, and the analytical proof that all *in silico* fingerprints satisfying the stated definition are found using our approach.
Results: The presented computational approach is implemented in an integrated high-performance computing (HPC) software tool for oligonucleotide fingerprint identification termed TOFI. We employed TOFI to identify *in silico* DNA fingerprints for several bacteria and plasmid sequences, which were then experimentally evaluated as potential probes for microarray-based diagnostic assays. Results and analysis of approximately 150 *in silico* DNA fingerprints for *Yersinia pestis* and 250 fingerprints for *Francisella tularensis* are presented.
Availability: The implemented algorithm is available upon request.
Contact: jaques.reifman@us.army.mil.

INTRODUCTION

The recent advances in genomic sequencing and the availability of large-scale sequence databases have unlocked several opportunities to identify 'genomic signatures' or 'DNA fingerprints', i.e. short DNA sequences that uniquely ascertain the presence or absence of causative biological agents, such as viruses, bacteria or virulent genes. For example, a vast number of DNA-based detection and diagnostic technologies are being developed to quickly identify biological threat agents (Ivnitski *et al*., 2003; Slezak *et al*., 2003; Draghici *et al*., 2005; Kaderali and Schliep, 2002), such as the anthrax-causing bacterium, *Bacillus anthracis*, and the plague-causing bacterium, *Yersinia pestis*. DNA signatures could also be used to detect the presence of one or more virulent genes, such as *Bacillus* genes, which encode important virulence factors, entereotoxins and exotoxins (Sergeev *et al*., 2006), and to provide high-resolution differentiation between closely related microorganisms in microbial forensics (Willse *et al*., 2004). New viruses and strains have been identified using a special microarray technology consisting of approximately 11 000 70mer oligonucleotides (Wang *et al*., 2002). DNA fingerprints have also been used to develop diagnostic assays for a wide-range of important applications in medicine, environmental monitoring and quality control of food products (Hardiman, 2003; Joos and Fortina, 2005; Wang *et al*., 2002; Abbe *et al*., 2004).

The specific algorithm implemented in a DNA fingerprint identification method is selected based on (1) whether the DNA fingerprints are being sought for a specific pathogen strain (e.g. *Y.pestis* CO92), a group of pathogens from the same species (e.g. all *Y.pestis* strains) or genus (e.g. all *Yersinia* species), or a set of organisms that may or may not have any phylogenetic relationship (e.g. to detect a viral or a bacterial family) and (2) the experimental conditions specified by the end application technology, such as PCR (Slezak *et al*., 2003; Viljoen *et al*., 2005; Haas *et al*., 2003; Gordon and Sensen, 2004) or DNA microarrays (Kaderali and Schliep, 2002; Hardiman, 2003; Rahmann, 2003; Leber *et al*., 2005; Nordberg, 2005).

The use of real-time PCR-based detection technology requires the identification of three informative sequences: two amplification primer sequences and an additional probe sequence (the fingerprint). The assay requires that primer hybridization takes place near the fingerprint and, therefore, imposes constraints on the position of the primer and PCR-based fingerprints. Moreover, PCR-based assays are quite limited in their multiplexing capabilities, as different assays are required to detect different pathogenic sequences. In contrast, microarrays do not impose any position specific constraints on the DNA fingerprints, and several fingerprints can be simultaneously placed on a microarray to provide detection redundancy and allow for the diagnosis of multiple pathogens on a single assay. Despite these advantages, microarray-based assays are relatively insensitive and slow compared to the exquisite sensitivity and speed of PCR-based assays. Microarray sensitivity can be greatly enhanced by incorporating sample amplification prior to hybridization but, unfortunately, this results in a net increase in assay time for already slow assays.

This paper is concerned with the identification of DNA fingerprints for specific, single pathogenic sequences, referred to as the

*To whom correspondence should be addressed.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 JAN 2007	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. Bioinformatics 23:3 - 13	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Tembe, W Zavaljevski, N Bode, E Chase, C Geyer, J Wasieloski, L Benson, G Reifman, J	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD	**PR-06-093**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

MOTIVATION: Advances in DNA microarray technology and computational methods have unlocked new opportunities to identify "DNA fingerprints," i.e., oligonucleotide sequences that uniquely identify a specific genome. We present an integrated approach for the computational identification of DNA fingerprints for design of microarray-based pathogen diagnostic assays. We provide a quantifiable definition of a DNA fingerprint stated both from a computational as well as an experimental point of view, and the analytical proof that all in silico fingerprints satisfying the stated definition are found using our approach. RESULTS: The presented computational approach is implemented in an integrated high-performance computing software tool for oligonucleotide fingerprint identification termed TOFI. We employed TOFI to identify in silico DNA fingerprints for several bacteria and plasmid sequences, which were then experimentally evaluated as potential probes for microarray-based diagnostic assays. Results and analysis of approximately 150 in silico DNA fingerprints for Yersinia pestis and 250 fingerprints for Francisella tularensis are presented. AVAILABILITY: The implemented algorithm is available upon request.

15. SUBJECT TERMS

methods, microarray, DNA fingerprint, identification, oligodeoxynucleotides, Yersina pestis, plague

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**9**	

target, for the design of DNA microarray-based diagnostic assays. The target could be an entire genome (e.g. *B.anthracis* Ames), a chromosome (e.g. *Brucella melitensis* biovar *abortus* 2308 chromosome II) or a non-chromosome sequence (e.g. *B.anthracis* plasmid pXO2). A more general problem, not addressed here, involves the identification of DNA fingerprints common to multiple strains or multiple species. An effective approach for this problem is to use multiple genome alignment and search for conserved regions to identify common DNA signatures (Slezak *et al*., 2003). Identification of common DNA signatures becomes an even greater problem for highly variable RNA viruses (Gardner *et al*., 2004), where a promising solution is to select combinations of non-unique probes and use unique hybridization patterns to unambiguously identify specific viral strains (Urisman *et al*., 2005; Schliep *et al*., 2003).

Given the long length of most targets, the identification of DNA fingerprints is a problem of high computational complexity. The potential solution space is extremely large because every subsequence of the target sequence needs to be considered. Furthermore, the determination of uniqueness requires comparison with nucleotide databases, such as the GenBank (Benson *et al*., 2005), that are growing exponentially in size. Moreover, uniqueness of DNA fingerprints obtained using such comparative algorithms is only valid with respect to the reference database used. As new sequences are made available, previously identified fingerprints need to be revalidated.

Several practical challenges arise in the experimental evaluation of the computationally identified DNA fingerprints for microarray-based assays. The DNA fingerprints should produce a high response when hybridized with a sample containing the target genome. Conversely, the response for any non-target genome should be as low as possible. Thus, algorithms for DNA fingerprint identification must include experimental constraints and DNA–DNA hybridization modeling methods to predict the response on a microarray. Although research in modeling molecular level interactions between DNA sequences has made significant progress, there is, unfortunately, no analytical method available today that can predict the exact outcome of a hybridization reaction between two or more arbitrary DNA sequences (SantaLucia and Hicks, 2004; Nordberg, 2005). Moreover, due to the variability in the outcome of a hybridization experiment, a large number of repetitions are required to experimentally evaluate the DNA fingerprints. It might not be possible to experimentally test all the computationally identified fingerprints because of the associated costs and limited resources.

To simultaneously accommodate these computational needs and experimental constraints, DNA fingerprint identification tools (Kaderali and Schliep, 2002; Rahmann, 2003; Leber *et al*., 2005) integrate computational algorithms for identifying unique sequences and DNA hybridization modeling tools for predicting the outcome of the microarray experiment. Often, these tools apply various approximations to reduce the computational complexity. For example, the integrated approach of Kaderali and Schliep (2002) uses an efficient search algorithm based on suffix trees and a simplified two-state transition near-neighbor thermodynamic model for DNA probe design and cross-hybridization. The simplified model reduces the computational time but introduces modeling errors in the DNA fingerprint design. A similar thermodynamic model with a computationally more efficient approach was proposed by Rahmann (2003). An efficient fractional programming approach for melting-temperature computation with an improved two-state transition near-neighbor thermodynamic model has also been proposed (Leber *et al*., 2005), however, computational time would still be an issue for cross-hybridization evaluation of a large number of non-target genomes.

Computational and experimental factors make quantification of the uniqueness or specificity of a short DNA sequence challenging. In fact, a literature survey indicates that the related studies have not stated a precise, quantitative definition of a DNA fingerprint. Although the general idea is to search the target genome for 'unique' DNA sequences and then test them experimentally, the *in silico* criterion for uniqueness has not been explicitly stated. In this paper, we first provide a formal definition of a DNA fingerprint based on various experimental conditions and a specificity criterion. We then describe an integrated approach that combines efficient bioinformatics algorithms, takes into account experimental constraints, and includes a large-scale comparison of DNA fingerprints with nucleotide databases. Next, we describe the algorithm underlying TOFI (tool for oligonucleotide fingerprint identification), its software implementation on a high-performance computing (HPC) platform, and an analytical approach to choose the input parameters of TOFI, which guarantees that all possible DNA fingerprints satisfying the stated definition are obtained. Finally, we discuss initial experimental results, which help evaluate our definition of a DNA fingerprint and the associated specificity criterion.

TERMINOLOGY AND PROBLEM DEFINITION

A DNA fingerprint for a given target genome g_t is defined with respect to a reference nucleotide sequence database denoted by $G = \{g_1, g_2, \ldots, g_n, \ldots, g_N\}$ that contains N sequences. In practice, G consists of DNA sequences from one or more publicly available comprehensive databases, such as GenBank, or any other smaller nucleotide database, such as a viral DNA sequence database. The target genome may or may not belong to G, implying that it could be a known pathogen or a newly sequenced one. Implicit in the definition of a DNA fingerprint is its validity with respect to the available database. As newer DNA sequences become available and are added to G, it is necessary to verify the validity of the previously identified fingerprints.

Based on the school of thought, computational or biological, the definition of a DNA fingerprint varies. Therefore, some discussion about our notion of a DNA fingerprint is in order.

From a pure computer science standpoint, a DNA fingerprint of g_t could be defined as 'any subsequence of g_t that is not a subsequence of any $g_n \in G$, $n \neq t$'. By this definition, the problem of identifying DNA fingerprints is equivalent to the classic string comparison problem of identifying substrings of g_t that do not exactly match any substring of any $g_n \in G$, $n \neq t$. Although mathematically correct, this definition lacks the application-specific requirements. DNA fingerprints have to satisfy: (1) design constraints, so that they can be used as DNA probes on microarrays and (2) specificity constraints, so that they can discriminate, in a microarray hybridization reaction, between target and non-target sequences. DNA fingerprints that simultaneously satisfy both design and specificity constraints require a biologically more sound definition. We mathematically formalize the experimental and specificity constraints as follows.

Let K denote the DNA microarray experimental constraints, such as the minimum and maximum length of the DNA fingerprint, the

hybridization melting-temperature, GC content, etc. (SantaLucia and Hicks, 2004), and let $P = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}$ denote the set of all subsequences of g_t that satisfy K. Thus, by definition, every $p_i \in P$ will have length within the specified minimum ($L_{\min}$) and maximum ($L_{\max}$) bounds, GC content within the required range, and will satisfy several other properties specified by the chosen DNA hybridization modeling methodology. We refer to the sequences in P as DNA probes. Note that constraints denoted by K do not specify attributes regarding the uniqueness of a DNA probe with respect to non-target sequences.

Quantifying specificity of DNA fingerprints from an experimental point of view is very subjective. It is based on interpreting the experimental hybridization results between DNA probes and non-target DNA sequences. Since there is a lack of accurate *in silico* hybridization models, we infer the specificity of a DNA probe by first computing DNA sequence alignments, and then determining if the aligned probe meets an empirical threshold T. This is based on the hypothesis that DNA sequences that align poorly are unlikely to form a stable DNA–DNA duplex for a given set of experimental constraints. This hypothesis implies that the DNA sequence alignment, calculated strictly using computational tools, provides quantification, through the threshold T, of the actual strength of the DNA–DNA duplex. Thus, we compute the specificity of a DNA probe from the number of mismatches, gaps and insertions/deletions in the alignment and compare it with the threshold T, representing the set U of all specificity constraints.

Having formalized the experimental and specificity constraints, we define a DNA fingerprint and the problem of identifying all DNA fingerprints for a target genome as follows:

Definition (DNA fingerprint): A DNA probe p_i of length L_i is considered a DNA fingerprint of g_t if and only if an optimal sequence alignment between p_i and any other sequence $g_n \in G$, $n \neq t$, has at most L_i-T matches.

Definition (DNA fingerprint identification): For a target DNA sequence g_t, find all *in silico* DNA fingerprints that satisfy the experimental constraints K and specificity constraints U with respect to a reference DNA sequence database G.

Let $S = \{s_1, s_2, \ldots, s_f, \ldots, s_F\}$ be a subset of P, i.e. $S \subseteq P$, that denotes the set of DNA probes that satisfy both constraints K and U. Our goal is to find all F elements of S. We refer to the elements of S as *in silico* DNA fingerprints because they satisfy all constraints that have been quantified for computational purposes. Their experimental validity needs to be tested in an actual DNA microarray experiment. Unless stated otherwise, henceforth the term 'DNA fingerprint' implies *in silico* DNA fingerprint, which is valid with respect to a reference database used.

INTEGRATED APPROACH

TOFI implements a multi-step approach that breaks down the problem of fingerprint identification into the three steps illustrated in Figure 1. The first step reduces the solution space by discarding DNA sequences common to both the target sequence and one or more biological near-neighbor sequences. The surviving sequences are termed candidate sequences. In the second step, a microarray DNA probe design phase extracts from candidate sequences only those subsequences that satisfy the application-specific experimental constraints K. In the third step, each DNA probe is aligned with all DNA sequences present in the chosen reference nucleotide database. A user-defined specificity constraint U is used to interpret the alignments from the standpoint of cross-hybridization between the DNA probes and non-target genomes. DNA probes satisfying U are reported as DNA fingerprints and are tested on microarrays.

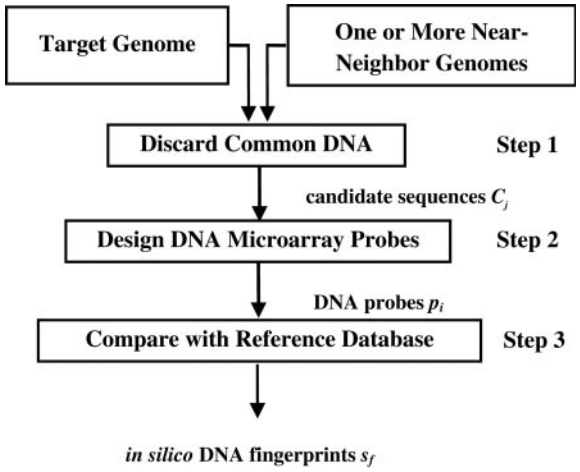

Fig. 1. The Three Steps of the TOFI Algorithm.

Although the problem has been split into three discrete steps for clarity of explanation, the individual steps are not completely independent from an algorithmic standpoint. In fact, our objective, to obtain all DNA fingerprints, leads us to conclude that the input parameters in the first step have an analytical relationship with the constraints imposed in the second and the third steps. We discuss the interdependence of the three steps after an in-depth description of each of the three steps.

Step 1: solution space reduction

The solution space to be searched is extremely large because every subsequence of the given target must be considered. Testing each subsequence experimentally is impractical and expensive. But reducing the solution space computationally can be quick and cheaper. For this purpose, we exploit the sequence similarities between the target genome and an evolutionary near-neighbor (g_r) that can be identified from a phylogenetic tree or published data. The target and neighbor will contain common DNA sequences, which, obviously, cannot be used as DNA fingerprints. DNA sequences common to both g_t and g_r are extracted using suffix trees (Wiener, 1973; Gusfield, 1997), which, within the domain of comparative genomics, have been used in Vmatch (Kurtz, 2002) and the Maximal Unique Matcher (MUMmer) (Kurtz *et al*., 2004) to identify repeats, exact or approximate matches, and single nucleotide polymorphisms. Details of the construction, traversal and numerous applications of suffix trees in several different string-matching applications are available in (Gusfield, 1997). It should be noted that the solution space could be further reduced by comparing the target genome with multiple non-target genomes, as described by Slezak and colleagues (Slezak *et al*., 2003). This could be done within TOFI's current algorithmic configuration by concatenating strings from multiple non-target genomes into one long string and providing it as the near-neighbor genome for comparison. An alternate, perhaps more efficient

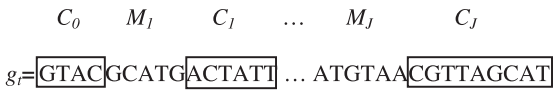

Fig. 2. Output of the Suffix-Tree-Based Algorithm.

approach, which would require software modification, is to compare the target sequentially against a list of non-target sequences, so that after each comparison only unmatched sequences are compared with the subsequent non-target genomes from the list.

Once the exact matches between the target g_t and its near-neighbor g_r are identified, the target can be represented as a concatenation $g_t = C_0M_1C_1M_2C_2 \ldots M_JC_J$, as shown in Figure 2. M_j denotes the jth exact match of length $|M_j|$ and C_j denotes the jth candidate sequence, i.e. a sequence in the target that contains no matches of length M or longer with the near-neighbor. C_0 and/or C_J can be null based on whether or not there is an exact match at the beginning or the end of g_t, respectively. Exact matches that are longer than the minimum length M are discarded and only the candidate sequences are retained for further consideration. The candidate sequences have no restriction with respect to their length, position in the genome, or composition of base pairs.

Of particular interest to our application is the choice of input parameters to the suffix-tree-based algorithm, in particular, the minimum length M of exact matches between g_t and g_r that would lead to identification of all DNA fingerprints. Our analysis indicates that the parameter M is closely related to the experimental and specificity constraints, detailed in Steps 2 and 3, respectively. Therefore, we first describe the remaining two steps of TOFI before an analytical relation is derived between the parameter M and the constraints imposed by the problem definition.

This selection of M differs from a related study (Slezak *et al*., 2003), where $M = 18$ was heuristically selected to meet the minimal PCR primer size.

Step 2: microarray probe design

The second step imposes a set of experimental constraints K to extract DNA microarray probes from the candidate sequences. A recent review (Panjkovich and Melo, 2005) indicates that for the same input DNA sequences different *in silico* probe design modeling tools, not surprisingly, produce different sets of DNA probes. To our knowledge, there is no universally accepted modeling methodology available today to design microarray probes from DNA sequences. Often, these tools are used in an iterative, trial-and-error fashion to optimize the quality/number of output DNA probes to suite the application-specific needs.

We have selected a probe design tool that implements a multi-state thermodynamic model for melting-temperature (SantaLucia and Hicks, 2004). The model allows for the representation of several dozen constraints on the DNA probes, such as probe length, GC content, molar concentrations, self-hybridization possibilities and limit on the number of single nucleotide repeats. Additional information on the constraints K can be found in SantaLucia and Hicks (2004). As an example, only a few important constraints are shown in Table 1. The DNA probes satisfying these constraints are extracted from every candidate sequence and are passed on to the next step.

Table 1. Typical values for DNA probe design constraints K

Length (bases)	Melting-temperature (°C)	GC content (%)
Minimum, L_{min}: 35 Maximum, L_{max}: 40	Minimum, T_{min}: 70 Maximum, T_{max}: 75	45–50

Step 3: specificity determination by sequence alignment

In the third step, every DNA probe is aligned with sequences in the reference nucleotide database. The results of the alignments are interpreted to predict cross-hybridization using the following general rule: A DNA probe that aligns poorly with all non-targets DNA sequences is unlikely to cross-hybridize with non-targets and, therefore, should be considered as a fingerprint.

Due to the limitations of DNA–DNA hybridization models, determining the alignment corresponding to the optimal DNA–DNA duplex on a microarray is hard. Computationally, optimal alignment between two DNA sequences could be defined using the generalized edit distance algorithm (Gusfield, 1997). Simply put, the edit distance between two sequences corresponds to the total number of insertions, deletions and substitutions that are needed to transform one sequence into the other. From the standpoint of DNA cross-hybridization, a substitution corresponds to a mismatched pair of nucleotides and insertions/deletions correspond to gaps in the DNA–DNA duplex. The lower the number of mismatches and gaps in the alignment is the lower in the edit distance. However, edit distance does not provide sufficient information with regards to the strength of hybridization. For example, it does not consider the position of matches in the alignment, GC content, gap length and the longest common factor in the alignment (Rahmann, 2003). Moreover, computing the optimal alignment between each DNA probe and every DNA sequence in a large database, such as the 'nt' nucleotide database from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) (Pruitt *et al*., 2005), would be very computationally intensive.

Based on these issues and practical time constraints, we opted to use the BLASTN program from BLAST (Altschul *et al*., 1990) for aligning DNA probes with a reference database. The alignment algorithm in BLAST is a heuristics-based approach that starts off by identifying a word of exact match of a given length (w parameter) and proceeds by extending it using dynamic programming to allow mismatches and gaps in the alignment. A statistical significance score termed E-value is used to distinguish between potentially meaningful alignments and chance alignments. The E-value score was used in (Draghici *et al*., 2005) to quantify the specificity of DNA probes. However, E-values are determined by the length of the alignment, size of the query, size of the total database and several other parameters that are not related to the ability of a probe to form a cross-hybrid with a non-target genome.

Instead of using E-value alone to determine probe specificity, TOFI examines the actual alignments reported by BLAST and determines the specificity of a probe by taking into account the number of matches, mismatches and gaps in the alignment independent of its statistical significance. These specificity constraints U form the basis for the empirical threshold T, used in the following

Table 2. Algorithm for candidate sequence selection using the suffix tree output

Variables:	Procedure:
g_t = Target genome g_r = Near-neighbor genome of g_t L_{min} = Minimum probe length L_{max} = Maximum probe length T = The specificity constraint, i.e. the combined minimum number of mismatches, insertions and deletions in the optimal alignment between a probe and any non-target sequence, defining a fingerprint M = Input parameter of the suffix-tree-based algorithm that specifies the minimum length of exact matches between g_t and g_r M_j = jth exact match between g_t and g_r, where $j = 1, \ldots, J$ C_j = jth candidate sequence, $j = 0, 1, \ldots, J$, defined as a subsequence of g_t that is: bounded on both sides by exact matches of length at least M, or located between the start of g_t and the first exact match of length at least M, or located between the last exact match of length at least M and the end of the genome g_t E = Length of the extension of candidate sequences into the adjacent exact match(es)	Input: L_{max}, L_{min}, T, g_t and g_r (1) P_{cand} = empty set (2) Let $E = L_{max} - T$ (3) Let $M = E + 1 = L_{max} - T + 1$ (4) Using suffix tree, identify exact matches $M_1, \ldots, M_J$ of length at least M between g_t and g_r, and candidates $C_0, C_1, \ldots, C_J$ from g_t (5) For each candidate sequence C_j from the suffix tree output, (A) If C_j is a prefix of g_t, then extend C_j by E bases to the right. Go to step D (B) If C_j is a suffix of g_t, then extend C_j by E bases to the left. Go to step D (C) Extend C_j on the left and the right by E bases (D) Add all subsequences of C_j satisfying the length constraints to P_{cand} Result: Every candidate DNA probe of g_t satisfying constraints L_{max}, L_{min}, and T will be in P_{cand}

hypothesis to infer hybridization patterns from 'optimal' BLAST alignments: If the best BLAST alignment between a DNA probe and a non-target genome has more than T mismatches or gaps, then the DNA probe will be considered as an *in silico* DNA fingerprint.

It is well documented that using BLAST for assessment of cross-hybridization of a probe with non-target genomes will result in some non-specific probes (Rahmann, 2003; Nordberg, 2005). If a word of length w is not found in a database sequence, the probe alignment with the sequence will be skipped resulting in potential missed cross-hybridization. In other situations, partial alignments with probes may result in underestimated cross-hybridization. Two promising approaches could be considered to improve probe specificity. First, additional filtering of the probes selected as fingerprints by BLAST could be performed to augment the hypothesis relating alignment to hybridization. In this case, additional information would be extracted from analyses of the alignments reported by BLAST, such as the maximum number of contiguous matches or the position of matches in the probe alignment, and used as rules to improve specificity constraints. Second, better alignment algorithms could be implemented as a post-processing step, which would incorporate hybridization thermodynamics into the alignment evaluation to take into account hybridization stability (Leber *et al*., 2005). However, it must be emphasized that the lack of an accurate model to directly relate a DNA sequence alignment with its corresponding DNA–DNA hybridization leaves the choice of probe specificity characterization as an open question.

TOFI parameter selection

The three steps in TOFI implement different bioinformatics algorithms, each carrying out a different task using its own set of input parameters. However, the minimum length of the exact matches M in the first step is analytically related to the length constraints L_{min} and L_{max} on the DNA probes in the second step and the specificity threshold T used in the third step. In this section, we mathematically derive this analytical relationship.

The problem of selecting an appropriate M value could be stated as follows: given the length constraints L_{min} and L_{max} on the length of the probes and the specificity threshold T, find a relationship between L_{min}, L_{max}, T and M, which guarantees that no valid DNA fingerprints are discarded.

Our approach is initiated by extending each candidate sequence C_j, $j = 0, 1, \ldots, J$, by E bases into each side of the neighboring exact match. This prevents the possible discarding of signatures that include the boundaries of C_j. From the extended candidate sequences we construct a candidate DNA probe set P_{cand}, which contains every sequence satisfying the length constraints L_{min} and L_{max}. Only those DNA probes in P_{cand} that satisfy the experimental constraints will be included in the probe set P for alignment with the reference dataset, i.e. $P \subseteq P_{cand}$.

We choose E such that $M > E$. This condition guarantees that the overlaps between two adjacent extended candidates, if any, will be limited to the exact match region separating the two candidates. It also sets a lower limit on M, $M = E + 1$. To ensure that any candidate DNA probe of length L_i having less than or equal to $L_i - T$ exact matches is not discarded, the extension length should be $E = L_i - T$. Thus, the extension length E is constrained by $L_{min} - T \leq E \leq L_{max} - T$. Substituting for $M = E + 1$ and making a conservative selection, we obtain $M = L_{max} - T + 1$. Such selection will, most likely, generate a candidate probe set P_{cand} that contains some false positives, i.e. probes that do not satisfy the specificity constraints. The majority of such non-specific probes will be discarded after the BLAST alignment inspection. However, some false positives will remain due to possible missed matches in BLAST, as described in the previous section. The details of the candidate probe selection algorithm are given in Table 2.

Finally, we prove that if the candidate sequences are obtained using $E = L_{max} - T$, with $M = E + 1$, then all DNA fingerprints is included in the set P_{cand}. However, as S denotes the set of DNA fingerprints for g_t, it will suffice to prove that such selection for M guarantees that $S \subseteq P_{cand}$.

Assertion 1. By construction, every subsequence of g_t containing an exact match of length smaller than M is included in an extended candidate sequence (steps 5.A–5.C in Table 2).

Assertion 2. From each extended candidate sequence, every subsequence satisfying the length constraints $L_{\min}$ and $L_{\max}$ is included in P_{cand} (step 5.D in Table 2).

Assertion 3. From assertions 1 and 2, none of the sequences in P_{cand} can contain an exact match of length M or greater.

Assertion 4. By definition, a DNA fingerprint $s_i \in S$ contains at most $L_i - T$ exact matches when aligned with any non-target genome. Thus, the length of the longest exact match between s_i and any other non-target genome is $L_i - T$. Since, $L_i - T \leq L_{\max} - T = E < M$, from assertions 1, 2 and 3, every $s_i \in S$ is included in P_{cand}. □

RESULTS

Software implementation

We used MUMmer (Kurtz *et al*., 2004), an open source software that implements a suffix-tree-based algorithm and provides several options for comparing genomic sequences. The microarray DNA probe design from candidate sequences was carried out using the commercial software oligonucleotide modeling platform (OMP) (available at http://www.dnasoftware.com), which implements a state-of-the-art hybridization model (SantaLucia and Hicks, 2004). The BLASTN program from NCBI_BLAST (version 2.2.10) was used for aligning more than 2.0 million nucleotide sequences stored in the 'nt' nucleotide database at the NCBI. The database has grown significantly since we obtained the results described in this paper and we have downloaded the latest version, containing more than 3.6 million sequences, for future runs of TOFI.

The entire software pipeline was initially implemented on a HPC environment at the Advanced Biomedical Computing Center (http://www-fbsc.ncifcrf.gov/) using the High Throughput Computing support from SGI® on an Altix cluster consisting of 64 × 1.5 GHz Itanium 2 processors running Red Hat® Linux with 64 GB of shared memory. Once candidate sequences are obtained, TOFI takes advantage of the parallel programming opportunities on HPC resources. The DNA probe design using OMP has been parallelized using OpenMP by scheduling DNA probe design for each candidate sequence on a separate processor. The execution of BLASTN, by far the most computationally intensive part of TOFI, is parallelized by assigning batches of DNA probes to separate processors. In addition, several application-specific software modules to process DNA sequences, to compile results of the intermediate stages for analysis, and to process outputs of various stages were implemented. This choice of resources and software is just one way to implement TOFI's integrated approach shown in Figure 1. A different choice of software for suffix tree, DNA probe design and sequence alignments could be used as well. However, our particular choice represents, arguably, some of the best tools available for each of the three steps.

TOFI has since been ported onto one of the U.S. Department of Defense Major Shared Resource Center's Linux clusters, consisting of 128 dual processor nodes on a distributed memory system, where deployment of mpiBLAST (Darling *et al*., 2003) and execution of OMP on separate processors is being tested. In the current cluster implementation, we use mpiBLAST with 32 processors running in parallel, which, again, consumes the bulk of the computing time.

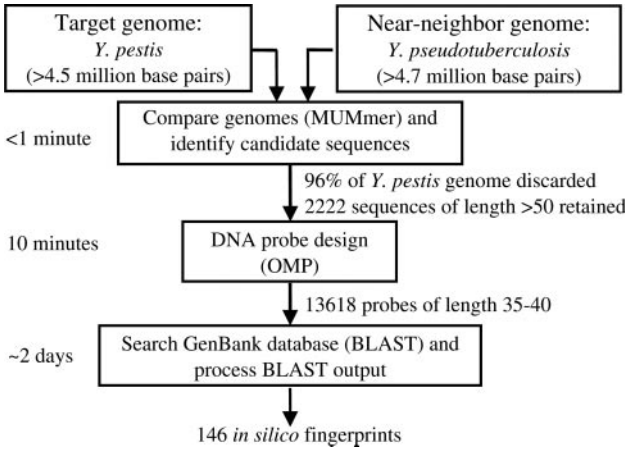

Fig. 3. Identification of *Y.pestis* DNA Fingerprints Using TOFI on a 32-CPU Linux Cluster.

The computational time of the algorithm depends on the number of probes generated as the output of Step 2 and provided to mpiBLAST and on the size of the reference database. The number of probes, in turn, depends on the length of the target genome, the availability and similarity of a near-neighbor genome, and the selected probe design constraints. The reference database is segmented according to rules of thumb suggested by the mpiBLAST developers (Darling *et al*., 2003), where the number of database segments is set to the number of processors. Hence, the computational time of processing the reference database is directly dependent on the speedup achieved by mpiBLAST (Darling *et al*., 2003). The execution time could be improved by using other parallel versions of BLAST, such as pioBLAST (Lin *et al*., 2005).

Case study: DNA fingerprints for *Y.pestis*

TOFI was used to identify DNA fingerprints for the plague-causing pathogen *Y.pestis* strain CO92 (accession no. NC_003143.1). Based on the literature (Chain *et al*., 2004), a closely related organism, *Yersinia pseudotuberculosis* strain IP 32953 (accession no. NC_006155.1), was selected as the near-neighbor genome.

The empirical threshold T=15 was selected based on a priori analysis of hybridization properties for the selected microarray technology and experimental setup, such as the probe length and required melting-temperature. This parameter can be adjusted on a case-by-case basis using feedback from additional experimental evaluations. For T=15 and maximum probe length $L_{\max} = 40$, according to step 3 in Table 2, the minimum length of exact matches in MUMmer is $M = 26$.

Approximately 96% of the *Y.pestis* genome was discarded using MUMmer in the first step (Fig. 3). Thus the idea of using a near-neighbor genome to identify and discard exact matches proved to be extremely effective in this case. Out of about 4.6 million bases of *Y.pestis*, fewer than 200 000 bases, distributed unevenly among 2222 candidate sequences, were considered further. In the next step, slightly over 13 600 DNA probes satisfying the experimental constraints were extracted from the candidate sequences.

In the BLAST probe specificity evaluation, the seed size $w = 7$ was selected because it was the smallest value available in the

BLAST version that we used, and a large E-value = 100 reduced the possibility of missing high scoring alignments.

Based on the specified TOFI parameters, all but 146 DNA probes were rejected, defining the *in silico* DNA fingerprints that were selected for experimental evaluation using custom DNA microarrays. These DNA fingerprints underwent further screening based on additional experimental constraints, such as the presence of restriction enzyme cleavage sites, leaving only 99 *in silico* fingerprints for testing.

Experimental evaluation of *in silico* fingerprints

Ten customized DNA microarray chips, each containing several replicates of the 99 *in silico* DNA fingerprints and a number of control sequences, were fabricated and used for evaluation purposes. Six chips were hybridized with the target genome *Y.pestis* and four chips were used to test cross-hybridization with the near-neighbor genome *Y.pseudotuberculosis*. Normalized data were used to compare hybridization signals.

The microarray hybridization data were used to analyze the discriminating power of the *in silico* fingerprints by comparing the experimental hybridization results of the probes with *Y.pestis* and *Y.pseudotuberculosis*. Figure 4 illustrates a sample set of data showing the normalized response (y-axis) as a function of the DNA fingerprints, which are arranged in descending order of the difference between their responses with *Y.pestis* and *Y.pseudotuberculosis*. Variability in the hybridization responses in repeated experiments is presented by standard error bars for each probe. Out of the 99 DNA fingerprints tested, 20 (data not shown in Fig. 4) produced higher average response for *Y.pseudotuberculosis* than that for the target. This is due to computational and experimental reasons. The computational reasons relate to limitations of using BLAST for specificity evaluation, as discussed in Section 3, step 3. A detailed post-experimental analysis of the BLAST outputs indicates that 12 out of the 20 probes do not have reported alignments with *Y.pseudotuberculosis* in the significant hit list. For the remaining eight probes, contiguous matches of 20 bases or more were observed in the BLAST alignments but the calculated sum of mismatches and gaps was larger than 15, causing these probes to be identified as fingerprints. This type of problem could be avoided if a threshold for the maximum number of contiguous matches could be experimentally determined and used for additional filtering of the probes that passed the first BLAST specificity testing. The experimental reasons relate to the variability of probe responses. Although all of the 20 probes have larger mean responses for *Y.pseudotuberculosis* than that for *Y.pestis*, only six of these probes have significantly larger responses. The experimental reason for the observed aberrant hybridization of these six probes is not clear. These 20 probes were excluded from further evaluation.

For a fingerprint to be useful in a diagnostic assay, it should yield a very low response for non-targets and a high response for the target. Thus, a few DNA fingerprints in Figure 4 that have a good discriminatory power but have a relatively high response for non-targets would not be considered useful on diagnostic assays. The data used in Figure 4 can also be used to identify valid fingerprints based on alternate rules, such as identifying quantifiable threshold values for target and non-target responses. For example, 25 probes could be selected by using a minimum threshold value of 2.0 for *Y.pestis* responses and a maximum threshold value of 1.0 for *Y.pseudotuberculosis* responses, while 20 probes could be selected using a minimum threshold of 2.0 for *Y.pestis* and allowing a maximum threshold of 0.5 for *Y.pseudotuberculosis*. In each case, a sufficiently large number of probes would allow for detection redundancy.

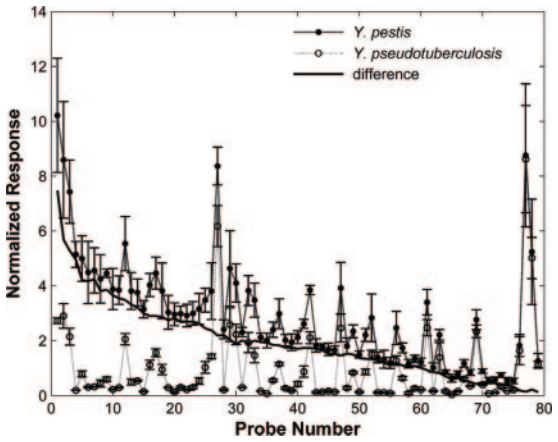

Fig. 4. Comparison of Hybridization of *in silico* Fingerprints with Target (*Y.pestis*) and Non-target (*Y.pseudotuberculosis*).

Other applications of TOFI

Having a near-neighbor genome is not a requirement for TOFI. The near-neighbor genome is used to reduce the solution search space as much as possible in the first step, which is computationally the least expensive step. The target genome could be compared with any small set of genomes using suffix trees. The higher the number of matches identified in the first step is, the lower is the number of computations required in the subsequent steps. In the current study, a single near-neighbor comparison reduced the search space very effectively. In the case in which a closely related near-neighbor for the target is unknown, either arbitrary genome(s) could be used as near-neighbor(s) or the first step could be omitted. In fact, TOFI was successfully used to identify DNA fingerprints for plasmids pPCP1, pCD1 and pMT1 in *Y.pestis* without using any near-neighbor. Because plasmids are much shorter (about a few thousand bases) than bacterial genomes (typically over a few million bases), the whole plasmid could be considered as a single candidate sequence and sent directly as input to the DNA probe design step.

TOFI was also used to identify fingerprints of *Francisella tularensis* strain SCHU S4 (accession no. NC_006570.1). The genomic sequence of the near-neighbor *Francisella philomiragia* was not available and, therefore, the first step of TOFI was omitted. In the second step, we used OMP to scan the whole *F.tularensis* genome, consisting of about 1.9 million bases. Overlap of adjacent probes was limited to 10 bases to reduce computation time. OMP identified about 20 000 probes, which were tested for specificity with $T = 15$, resulting in 250 fingerprints. Further screening for restriction enzyme cleavage sites reduced the number of *in silico* fingerprints to 121.

Four chips were fabricated using several replicas of the 121 *in silico* fingerprints and a number of control probes. Two chips were used to test hybridization with *F.tularensis* and the other two to test cross-hybridization with *F.philomiragia*. We performed

initial evaluation using a criterion similar to the one employed for *Y.pestis*. In contrast to the *Y.pestis* hybridization experiments, only one probe had higher average response with *F.philomiragia* than with *F.tularensis*. In the experiment, 85 probes showed a normalized response with *F.philomiragia* smaller than 1.0, while 81 of those probes had responses larger than 2.0 with *F.tularensis*.

Currently, a large number of additional experiments, including a standard panel of non-target genomes, are being performed to evaluate the fingerprints of *Y.pestis* and *F.tularensis* before they are used as probes in diagnostic assays.

Current limitations and plans for improvements

TOFI has already been used in its current configuration to identify fingerprints for a number of pathogens. However, several algorithmic and implementation issues affect its performance and are being addressed.

The scope of DNA fingerprint identification in TOFI is currently limited to a single target sequence. We are investigating approaches to select fingerprints common to a large number of related targets, which would allow for the identification of fingerprints common to specific species or genus. For highly variable RNA viruses, unique fingerprints may not exist. For this application, an approach based on the selection of non-unique probes, which together may form unique hybridization patterns on a chip for unambiguous viral identification, is also being considered.

Although we have provided a definition of a DNA fingerprint and an algorithm that guarantees all fingerprints satisfying it are identified, valid probes could potentially be discarded due to significant overlap with adjacent probes during the probe design phase (Step 2 of TOFI). This relates to practical considerations in order to reduce the number of 'similar' probes on the chip, allow space for multiple replicates, and limit the total number of probes, considering that several control sequences need to be present on the microarray.

Experimental evaluations of the identified *in silico* fingerprints for *Y.pestis* and *F.tularensis* indicate the possibility for improvement in the algorithm specificity. It was found that cross-hybridization with non-target genomes was not detected by BLAST in about 10% of the *in silico* fingerprints for *Y.pestis*, while in an additional 10% of the fingerprints the cross-hybridization was underestimated. Specificity will be improved in the future based on: (1) the selection of optimal TOFI parameters using comprehensive evaluation of the experimental results; (2) the post-processing of BLAST alignments using expert rules to better correlate alignments with hybridization and (3) the development of optimal alignment algorithms that include hybridization thermodynamics as a post-processing step after the BLAST specificity evaluation.

Due to the extremely rapid growth/modifications in available DNA sequences, the continued validity of the DNA fingerprints must be frequently verified. Efforts to automatically update fingerprints are also planned.

CONCLUSIONS

This work presented TOFI, an integrated bioinformatics tool to identify *in silico* genomic fingerprints for the design of microarray diagnostic assays. TOFI is a standalone application that exploits the parallel programming benefits provided by HPC platforms and allows users to select input parameters through a graphical user interface. This work differs from previous ones in that a formal definition of a DNA fingerprint is provided. More importantly, given the desired length of a fingerprint and its required number of non-matching base pairs, we provide an algorithm that guarantees that all *in silico* fingerprints are identified. Fingerprints for a number of pathogenic sequences have been preliminarily evaluated through experimental tests with pathogens of interest and non-target genomes. Initial results indicate that the approach is capable of identifying multiple fingerprints for specific DNA sequences, and that the algorithm could be improved to enhance specificity. Further testing, with a standard panel of non-target genomes, is underway. This information will enable optimal TOFI parameter selection and will serve as a valuable benchmark for future algorithm improvements.

DISCLAIMER

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or the U.S. Department of Defense.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to the anonymous referees for useful comments and suggestions as well as interesting ideas for future work. The authors thank Kamal Kumar of the Biotechnology HPC Software Applications Institute for help in developing TOFI's graphical user interface and Bob Stephens, Jack Collins and Karol Miaskiewicz of the Advanced Biomedical Computing Center, National Cancer Institute, Frederick, MD, for the computational support. This work was sponsored by the U.S. Department of Defense High-Performance Computing Modernization Program (HPCMP), under the High-Performance Computing Software Applications Institutes (HSAI) initiative, and the U.S. Defense Threat Reduction Agency.

Conflict of Interest: none declared.

REFERENCES

Abee,T. *et al*. (2004) Impact of genomics on microbial food safety. *Trends Biotechnol*., **22**, 653–660.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Benson,D.A. *et al*. (2005) GenBank. *Nucleic Acids Res*., **34**, D16–D20.

Chain,P.S. *et al*. (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **101**, 13826–13831.

Darling,A. *et al*. (2003) The design, implementation, and evaluation of mpiBLAST. In *4th International Conference on Linux Clusters: The HPC Revolution 2003*, *in conjunction with the ClusterWorld Conference & Expo,* San Jose, CA.

Draghici,S. *et al*. (2005) Identification of genomic signatures for the design of assays for the detection and monitoring of anthrax threats. In Altman,R.B. *et al*. (ed.), *Proceedings of the Pacific Symposium of Biocomputing 2005* Hawaii, USA, pp. 248–259.

Gardner,S. *et al*. (2004) Sequencing needs for viral diagnostics. *J. Clin. Microbiol*., **42**, 5472–5476.

Gordon,P.M.K. and Sensen,C.W. (2004) Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays. *Nucleic Acids Res*., **32**, e133.

Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK.

Hardiman,G. (ed.) (2003) *Microarray Methods and Applications-Nuts and Bolts*. DNA Press, Eagleville, PA, USA.

Hass,S.A. *et al*. (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res*., **31**, 5576–5581.

Ivnitski,D. *et al*. (2003) Nucleic acid approaches for detection and identification of biological warfare and infectious disease agents. *Biotechniques*, **35**, 862–869.

Joos,T. and Fortina,P. (2005) *Microarrays in clinical diagnosis*. Humana Press, Totowa, NJ, USA.

Kaderali,L. and Schliep,A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18**, 1340–1349.

Kurtz,S. (2002) construction and application of virtual suffix trees.. PhD dissertation, Technische Fakultöen, Universitat Bielefeld, Bielefeld, Germany.

Kurtz,S. *et al*. (2004) Versatile and open software for comparing large genomes. *BMC Genome Biol*., **5**, R12.

Leber,M. *et al*. (2005) A fractional programming approach to efficient DNA melting temperature calculation. *Bioinformatics*, **21**, 2375–2382.

Lin,H. *et al*. (2005) Efficient data access for parallel BLAST. *IEEE International Parallel and Distributed Processing Symposium,* Denver, CO.

Nordberg,E. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–1370.

Panjkovich,A. and Melo,F. (2005) Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics*, **21**, 711–722.

Pruitt,K.D. *et al*. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.

Rahmann,S. (2003) Fast large scale oligonucleotide selection using the longest common factor approach. *J. Bioinfo. Compu. Biol*., **1**, 343–361.

SantaLucia,J.,Jr and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct*., **33**, 415–440.

Schliep,A. *et al*. (2003) Group testing with DNA chips: generating designs and decoding experiments. In *Proceedings of the Computational Systems Bioinformatics, August 11-14,* Stanford, CA, pp. 84–91.

Sergeev,N. *et al*. (2006) Microarray analysis of *Bacillus cereus* group virulence factors. *J. Microbiol. Meth*., **65**, 488–502.

Slezak,T. *et al*. (2003) Comparative genomics tools applied to bioterrorism defense. *Brief. Bioinform*., **4**, 133–149.

Urisman,A. *et al*. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *BMC Genome Biol*., **6**, R78.

Viljoen,G.J. *et al*. (eds) (2005) *Molecular Diagnostics PCR Handbook*. Springer Publishers, Berlin, Germany.

Wang,D. *et al*. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc. Natl Acad. Sci. USA*, **99**, 15687–15692.

Weiner,P. (1973) Linear pattern matching algorithms. In *Proceedings of 14th IEEE Annual Symposium on Switching and Automata Theory,* Washington, DC, IEEE Computer Soc., pp. 1–11.

Willse,A. *et al*. (2004) Quantitative oligonucleotide microarray fingerprinting of *Salmonella enterica* isolates. *Nucleic Acids Res*., **32**, 1848–1856.